DOCUMENT RESUME

ED 219 425 TM 820 439

AUTHOR White, Karl R.; Carcelli, Larry
TITLE The Effect of Item Format on Students' Standardized Mathematics Achievement Test Scores.
PUB DATE Mar 82
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Analysis; Computation; Elementary School Mathematics; Grade 2; *Objective Tests; Primary Education; Quantitative Tests; Scores; *Standardized Tests; Test Construction; *Test Format; Test Validity
IDENTIFIERS Iowa Tests of Basic Skills; Keymath Diagnostic Arithmetic Test; SRA Achievement Series; Stanford Achievement Tests; Wide Range Achievement Test

ABSTRACT

The effect on children's test scores of different item formats used in standardized mathematics achievement tests was investigated. In a Utah school, 40 second grade students completed a mathematics computation test using eight different formats derived from five standardized achievement tests. Identical content, taken from the most frequently occurring content in the standardized achievement tests examined, was tested with each format. Differences in test score between types of formats were statistically significant. These results indicate that what a student appears to have mastered in mathematics computation is substantially influenced by the format of the particular test used. Differences in scores between the various formats are not accounted for by the normative scaling of the different tests. Greater attention should be given to the effect of test item format in selecting and administering achievement tests. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219425

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Effect of Item Format on Students' Standardized Mathematics Achievement Test Scores

Karl R. White and Larry Carcelli

Utah State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. R. White

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper to be presented at the American Educational Research Association's Annual Meeting in New York

March, 1982

TM 820 439

Abstract

The effect on children's test scores of different item formats used in standardized mathematics achievement tests was investigated. Second grade students completed a mathematics computation test using eight different formats derived from five standardized achievement tests. Identical content, taken from the most frequently occurring content in the standardized achievement tests examined, was tested with each format. Differences in test score between types of formats were statistically significant at $p<.001$ ($F = 45.25$). These results indicate that what a student appears to have mastered in mathematics computation is substantially influenced by the format of the particular test used. Differences in scores between the various formats are not accounted for by the normative scaling of the different tests. Greater attention should be given to the effect of test item format in selecting and administering achievement tests.

The Effect of Item Format on Students' Standardized Mathematics Achievement Test Scores

Cronbach (1971) defined the universe of behavior sampled by a test as including all abilities required to perceive the stimulus items and formulate responses. For an achievement test to be a valid indicator of what a student knows, the student must get answers correct or incorrect because of what he or she knows. If variables besides the student's knowledge of the content being tested affect the student's score, correct interpretation of the results is more difficult, and one must question the validity of the test for making decisions that depend on the student's knowledge (e.g., educational placement or instructional programming decisions). For standardized achievement tests, one such factor that might influence test scores and confuse the interpretation of what a student appears to know is the format in which the test is administered.

Previous Research

Previous research has investigated a number of factors which may affect how well a student does on standardized achievement tests in addition to what the student knows. Whether test items are presented from easiest to most difficult, most difficult to easiest, or in scrambled order affects how difficult students perceive the test to be (Barciskowski & Olsen, 1975) and how anxious the test makes students feel (Towle & Merrill, 1975). Some researchers have reported that increasing item difficulty elicits better performance (Holliday &

Partridge, 1979), while others have reported no effect (Dambrot, 1980; Gerow, 1980).

Millman and Setijadi (1966) showed that Indonesian students perform better than American students on open-ended math problems while American students perform better on multiple-choice problems. They concluded that the format in which questions were asked was a determinant of students' scores. Other researchers have investigated the use of pictures vs. no pictures, three dimensional objects vs. pictures, and multiple choice vs. true-false vs. matching and have reached similar conclusions (Poage & Poage, 1977; Kierscht & Vietze, 1975; Kursan, Rabinsky, & Pandley, 1979; Loftus, 1971).

Kintsch (1968) suggested that recall items require students to both search for and retrieve information, whereas recognition items only require students to discriminate between the presented information. Frisbie (1973) hypothesized that "multiple choice questions limit the universe of comparisons" because they only require an individual to recognize the correct response. In answering a true-false question, however, an individual must eliminate the incorrect response by forming a counter example from a wide universe of possibilities.

Variations in the answer format may also affect the student's score on an achievement test. Grier (1975), Weber (1977), and Catts (1978) found that scores were lower when additional distractors were used. Williams, Davis, Anderson, and Favor (1978) and Lai-Min and

Coffman (1974) found that an additional "all incorrect" distractor used with the Iowa Test of Basic Skills decreased scores.

Results from achievement tests are usually interpreted as an indication of what children know and are frequently used to make important decisions concerning the placement, advancement, and educational programming of children. Based on previous research it appears that how the test items are organized and the way test items are asked and answered all affect students' performance on tests. However, the issue of how item format affects what students appear to know has not been sufficiently investigated. Previously reported research has not explicitly investigated the impact on student test scores of differences in item format based on format differences found in frequently used standardized achievement tests. Since the results from standardized achievement tests are used extensively to estimate what students know, it is important to determine whether scores for the same content would be different depending on which format is used.

Method

Test Construction

The mathematics computation subtests of five standardized achievement tests were analyzed, and eight different types of question formats for testing computational skills were identified as shown in Table 1.

Insert Table 1 about here

The content tested by the computational subtests of the same standardized tests was analyzed, and 14 different types of addition and subtraction problems were identified. The four most commonly occurring content types were:

- Addition of two single digit numbers without carrying (e.g., 4 + 5 = 9).
- Addition of a single and a double digit number with carrying (e.g., 4 + 17 = 21).
- Subtraction of a single digit number from a single digit number without borrowing (e.g., 8 - 3 = 5).
- Subtraction of a single digit number from a double digit number with borrowing (e.g., 15 - 8 = 7).

A Format Test of 96 questions was constructed so that the same 12 addition or subtraction problems were tested in each of the eight format types. Three questions from each of the four most frequently occurring content types were generated and randomly arranged in each format group to make up the 12 items. The content for the 12 items is shown in Table 2.

Insert Table 2 about here

Pilot Studies

The Format Test was administered to 10 second grade elementary school children during two pilot studies. The following problems with the original version of the Format Test were discovered and corrected. First, some of the original items were too easy, resulting in a ceiling effect. Consequently, the two easiest content types (addition of single digit numbers without carrying and subtraction of single digit numbers without borrowing) were deleted and the next most frequently occurring, but more difficult, content types (double/single digit

addition without carrying and double/double digit subtraction without borrowing) were added. Furthermore, two additional items were added to formats 2, 3, 5, and 6 which contained "none-of-the-above" (N) distractors. The correct answer to these added items was N to suggest to the students that the N distractor was a viable alternative. Thus, the final form of the Format Test contained 104 items, of which only 96 were analyzed.

Reliability of the Format Test

Item analyses and Hoyt ANOVA estimates of reliability were computed for the 12 items in each format type. Only eight of the 96 items had point biserial correlations below r = +.20, four of which were attributable to the extreme difficulty levels of those particular items. The Hoyt estimate of reliability for the entire test (96 items) was .95 with the reliability of the 12 item tests for each of the eight formats ranging from .67 to .84.

Procedures

The Format Test was administered to 40 second grade students (18 male and 22 female) from two classrooms (20 students in each classroom) in a single school in Utah. The two classes were arbitrarily selected by school district personnel. All students were Caucasian and from a semi-rural area.

The Format Test was composed of eight sections or format groups that tested identical content. To investigate the presence of and counteract a possible practice effect resulting from identical content being tested in eight different formats, the Format Test was divided in

half (i.e., four format groups per half) and separately administered in two parts to the two classes during four testing sessions. The order of administration of the two halves of the test was varied between the two classes and counterbalanced across sessions.

Prior to the administration of the 12 items in each format, two sample test items of the type to be tested were computed on a blackboard with the help of the children. Instructions from the achievement test from which the format type was derived were then given. The children had ample time (12 minutes) to complete each item in all format groups.

Results

Format Tests

The cell and marginal means and standard deviations of the percentage of items the students answered correctly (the dependent variable) in each content category and format group (the independent variables) are presented in Table 3. For example, the students answered 73.3% of the items correctly when the content category of double/double digit addition without carrying was tested with Format 4. The average percentage of correct responses across all four content categories when tested with Format 4 was 81.7%. As can be seen in Table 3, the largest difference in the percentage of items answered correctly is between Formats 8 (33.8%) and 4 (81.7%), for a difference of 47.9%.

Insert Table 3 about here

A three-way repeated measures ANOVA indicated a statistically significant difference in the percent of items answered correctly by the students in each of the format groups (F = 45.25, $p<.001$) and content types (F = 17.08, $p<.001$), and an interaction between format groups and content types (F = 12.27, $p<.001$) (see Table 4). The educational or practical significance of the differences between format types is indicated by the fact that the difference between Formats 4 and 8 represent almost 3 standard deviations on the Format Test. Furthermore, the ω^2 (omega squared, as described in Hays, 1973) associated with the Between Format Sums of Squares is .183.

Insert Table 4 about here

Scheffé Analyses

Because the omnibus test for the format main effect was statistically significant, multiple comparisons were computed using the Scheffé technique to gain further information about the influence of specific test item characteristics on students' scores. The specific formats compared and the results of the Scheffé analyses are shown in Table 5. Statistically significant differences between pairs ($p<.01$) were found for all pairwise comparisons except the horizontal vs. vertical and the verbal vs. visual formats.

Insert Table 5 about here

Counterbalance

The cell and marginal means and standard deviations of the percentage of items the students answered correctly in each class and session are presented in Table 6. The percentages are nearly equal across both classes and sessions. The greatest difference is only 2.1% between session #1 (59.6%) and #2 (61.7%). A three-way repeated

Insert Table 6 about here

measures ANOVA using session order, class, and replicates as the independent variables and students' scores on the Format Test as the dependent variable did not indicate statistically significant main effects for sessions ($F = .72$, $p<.25$), classes ($F = .21$, $p<.25$), or the interaction between sessions and classes ($F = .18$, $p<.25$) (see Table 7).

Insert Table 7 about here

Discussion

The results of this study indicate that children score very differently on mathematics computation items depending on the format in which the test items are presented. This means that conclusions about how well children have mastered mathematical content will depend in part upon the format of the particular standardized test which is used. The differences in percentage correct scores between formats were not due to a practice effect. The children did not score better on the format groups that were administered near the end of the testing

session or during the last test session as a result of repeated exposure to identical content.

Specific characteristics of the item formats affected the scores children received on the Format Test. Scheffé analysis #1 (i.e., Box vs. Vertical) and #2 (i.e., Box vs. Horizontal) showed that Box problems were more difficult than other forms of question presentation. These results indicate that in addition to math computation skills, the Box problems tested complex concept level skills that required the students to understand the reversibility of signs.

Scheffé analysis #4 indicated that verbal and visual presentation of items did not differentially affect scores. However, when the two box problems were deleted from the analysis (because they were both visually administered), Scheffe analysis #7 indicated that the students obtained significantly lower scores on problems that were verbally administered. It may be that children received lower test scores with verbally administered problems because the test stimuli were transient and required the children to remember the type of problem and the value of the numbers being computed while at the same time determining the correct answer.

Children scored higher on items that used multiple-choice response formats than on items that used open-ended response formats (Scheffé analysis #5). In support of past research these differences can probably be attributed to the different abilities required to correctly answer a problem in the two formats. With a multiple-choice format, children needed only recognize the correct response and mark it.

However, with the open-ended format, children needed to retrieve the correct response from memory before answering. Apparently, retrieving a correct answer from memory is more difficult than simply recognizing the correct answer.

Children scored higher on multiple-choice response formats without "none-of-the-above" options than on multiple-choice response formats with "none-of-the-above" options (Scheffé analysis #6). Multiple-choice response formats with "none-of-the-above" options, like open-ended response formats, may not have visually presented the correct answer to the children; whereas the correct answer was visually presented to the children with multiple-choice response formats without "none-of-the-above" options.

Since raw score percentage correct was used as the dependent variable in the reported analyses, differences between formats would be less important if they were adjusted for in the test's norms. In other words, a difference between a score of 40% correct using a format from Test A and 80% using the format of Test B would not be as important if the 60th percentile for Test A was 40% correct and the 60th percentile for Test B was 80% correct. In Figure 1 the mean percentage of students scoring correct for each test format has been transformed to percentile scores using the norms from the various standardized test publishers (such transformations were not possible for Formats 1, 7, and 8). As can be seen, the differences between the various types of formats are as great or greater when percentiles are used instead of percentage correct scores.

Insert Figure 1 about here

The actions resulting from interpretation of children's standardized achievement test scores can have a significant influence on children's educational experiences. If format variables are responsible for substantial variation in students' scores, then it is very difficult to interpret the children's scores correctly and to take appropriate educational actions based on those scores. School personnel should, therefore, pay as much attention to the format used by a test as they have traditionally paid to the content of the test when selecting and interpreting children's scores on standardized achievement tests. This is particularly true when using children's scores on individual test items as criteria for content mastery, because children may answer individual items incorrectly for reasons other than lack of knowledge.

Additionally, in designing achievement tests, test constructors should more thoroughly assess the skills required to understand and respond to the format used by the test. For example, the abilities required by items which use a box format may more appropriately be assessed in a mathematical concept subtest. At the least, the abilities required by the test formats should be more specifically defined and communicated in the test manuals.

One explanation for the observed differences is that students may perform better on those tests which ask questions in the same way that is familiar to the student. In other words, using an unfamiliar format

may well depress students' scores by introducing an irrelevant variable and make it appear that a student has not mastered the material as well as he or she has. Future research needs to investigate more systematically the degree to which students' familiarity with a particular format may be responsible for differences between the students' scores. If degree of familiarity with the format that is used in a standardized test is responsible for much of the variance that has been reported, then educators should make sure their students are thoroughly familiar with the formats in a test before using the results of that test to make important educational decisions.

References

Barciskowski, R. S., & Olsen, H. Test item arrangement and adaptation level. The Journal of Psychology, 1975, 90, 87-93.

Catts, R. How many options should a multiple choice question have? (a) 2. (b) 3. (c) 4. At a glance research report. Sydney, Australia: New South Wales Department of Education, 1978. (ERIC Document Reproduction Service No. ED 173 354)

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Dambrot, F. Test item order and academic ability, or should you shuffle the test item deck? Teaching of Psychology, 1980, 7, 94-96.

Frisbie, D. A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Gerow, J. R. Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 1980, 7, 93-94.

Grier, L. B. The number of alternatives for optimum test reliability. Journal of Evaluation Measurement, 1975, 12, 109-112.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart, and Winston, 1973.

Holliday, W. G., & Partridge, L. A. Differential sequencing effects of test items on children. Journal of Research in Science Teaching, 1979, 16, 407-411.

Kierscht, M. S., & Vietze, P. M. Test stimuli: Representational level with middle class and Head Start children. Psychology in the Schools, 1975, 12, 309-312.

Kintsch, W. Recognition and free recall of organized lists. Journal of Experimental Psychology, 1968, 76, 481-487.

Kursan, V. K., Rabinsky, L., & Pandley, T. N. Test mode, test instructions, and retention. Contemporary Educational Psychology, 1979, 4, 211-218.

Lai-Min, P. S., & Coffman, W. E. A study of "I don't know" response in multiple-choice tests. Iowa City, Iowa: Iowa Testing Programs, University of Iowa, 1974. (ERIC Document Reproduction Service No. ED 141 371)

Loftus, G. R. Comparison of recognition and recall in a continuous memory task. Journal of Experimental Psychology, 1971, 91, 220-226.

Millman, J., & Setijadi. A comparison of American and Indonesian students on three types of test items. Journal of Educational Research, 1966, 59, 273-275.

Poage, M., & Poage, E. G. Is one picture worth a thousand words? Arithmetic Teacher, 1977, 24, 408-414.

Towle, N. J., & Merrill, P. F. Effects of anxiety type and item-difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 1975, 112, 241-249.

Weber, M. B. The effect of choice format on internal consistency. Emory University, 1977. (ERIC Document Reproduction Service No. ED 161 940)

Williams, R. L., Davis, W., Anderson, P., & Favor, K. Test format as a form of bias for black students. Journal of Non-white Concerns in Personnel and Guidance, 1978, 6, 141-147.

Table 1

Format Differences in Math Subtests of

Standardized Achievement Tests

	TYPE OF FORMAT				
FORMAT #1:	Wide Range Achievement Test — Level I "Jack had six marbles. He found two more. How many did he have?" Oral Response				
FORMAT #2:	Stanford Achievement Test — Level II — Form A "Read the problem to yourself and then mark under your answer." $2 + \square = 8$	11 0	6 0	5 0	N 0
FORMAT #3:	Stanford Achievement Test $\begin{array}{r} 2 \\ +\,6 \\ \hline \end{array}$	26 0	8 0	5 0	N 0
FORMAT #4:	SRA Achievement Series — Level C — Form 1 "Work the problem, then fill in the space in front of the right answer." $\boxed{\begin{array}{r} 2 \\ +\,6 \\ \hline \end{array}}$	8 0	4 0	26 0	12 0
FORMAT #5:	Iowa Test of Basic Skills — Level 7 — Form 7 "What is two added to six? What is two plus six? Mark your answer in the row with the bird." 	7 0	8 0	9 0	N 0
FORMAT #6:	Iowa Test of Basic Skills $6 + 2 =$	12 0	8 0	10 0	N 0
FORMAT #7:	Key Math "These are addition problems. You may use a pencil if you wish." $\begin{array}{r} 6 \\ +\,2 \\ \hline \end{array}$				
FORMAT #8:	Key Math "Tell me the number that goes in the box." $6 + \square = 8$				

Table 2

Content Used for the 12 Items in Each of the Eight Formats in the 96 Item-Format Test

14 + 9	28 -12	16 +22	23 +14	31 - 5	32 - 7
29 -15	19 -17	26 -11	16 + 5	33 - 9	15 + 7

Table 3

Cell and Marginal Means and Standard Deviations of Percentage Correct Scores for Format and Content Type on the Format Test

Test Format			Content Type 2b Double/single digit with carrying	3a Double/double digit without carrying	6b Double/single digit with borrowing	7a Double/double digit without borrowing
			.658 (.050)	.690 (.088)	.516 (.092)	.569 (.039)
WRAT	1	.496[a] (.112)[b]	.716 (.139)	.483 (.001)	.350 (.053)	.433 (.113)
SAT	2	.383 (.225)	.416 (.057)	.200 (.144)	.450 (.158)	.475 (.041)
SAT	3	.731 (.123)	.850 (.037)	.733 (.041)	.550 (.089)	.792 (.010)
SRA	4	.817 (.209)	.850 (.048)	.733 (.106)	.550 (.066)	.792 (.008)
ITBS	5	.677 (.069)	.900 (.141)	.708 (.071)	.492 (.093)	.608 (.119)
ITBS	6	.685 (.077)	.825 (.058)	.633 (.013)	.642 (.049)	.642 (.093)
Key Math	7	.738 (.130)	.767 (.052)	.775 (.008)	.483 (.162)	.925 (.137)
Key Math	8	.338 (.271)	.200 (.219)	.133 (.165)	.500 (.255)	.517 (.129)

Note. Examples of format types are shown in Table 1. The overall mean = .583, and the overall standard deviation = .171.

[a]Mean.

[b]Standard deviation.

Table 4

Sources of Variance for Repeated

Measures ANOVA for Format Test

Source	Degrees of freedom	Error term	Sum of squares	F
Format (F)	7	FR	320.39	45.25*
Content (C)	3	CR	55.48	17.08*
Replicates (R)	39		352.17	
F X C	273	FCR	139.90	12.27*
F X R	273		276.14	
C X R	117		126.66	
F X C X R	819		444.69	

*$p < .001$.

Table 5

Scheffe Pairwise Comparisons of Different Formats

Description of comparisons	Formats compared	F scores
1. Box vs. vertical	2,8 vs. 3,4,7	745.57*
2. Box vs. horizontal	2,8 vs. 6	100.30*
3. Horizontal vs. vertical	6 vs. 3,4,7	6.23
4. Verbal vs. visual	1,5 vs. 2,3,4,6,7,8	2.09
5. Open vs. multiple-choice	1,7,8 vs. 2,3,4,5,6	47.90*
6. No (N) vs. (N)	4 vs. 2,3,5,6	44.07*
7. Verbal vs. visual - box	1,5 vs. 3,4,6,7	46.36*

*$p < .01$.

Table 6

Cell and Marginal Means and Standard Deviations of Percentage Correct Scores for Session and Class for Counterbalance

Class		Session	
		Session #1	Session #2
		.596[a]	.617
		(.191)[b]	(.186)
Class #1	.619	.578	.602
	(.216)	(.173)	(.193)
Class #2	.591	.601	.638
	(.193)	(.208)	(.179)

[a]Means.

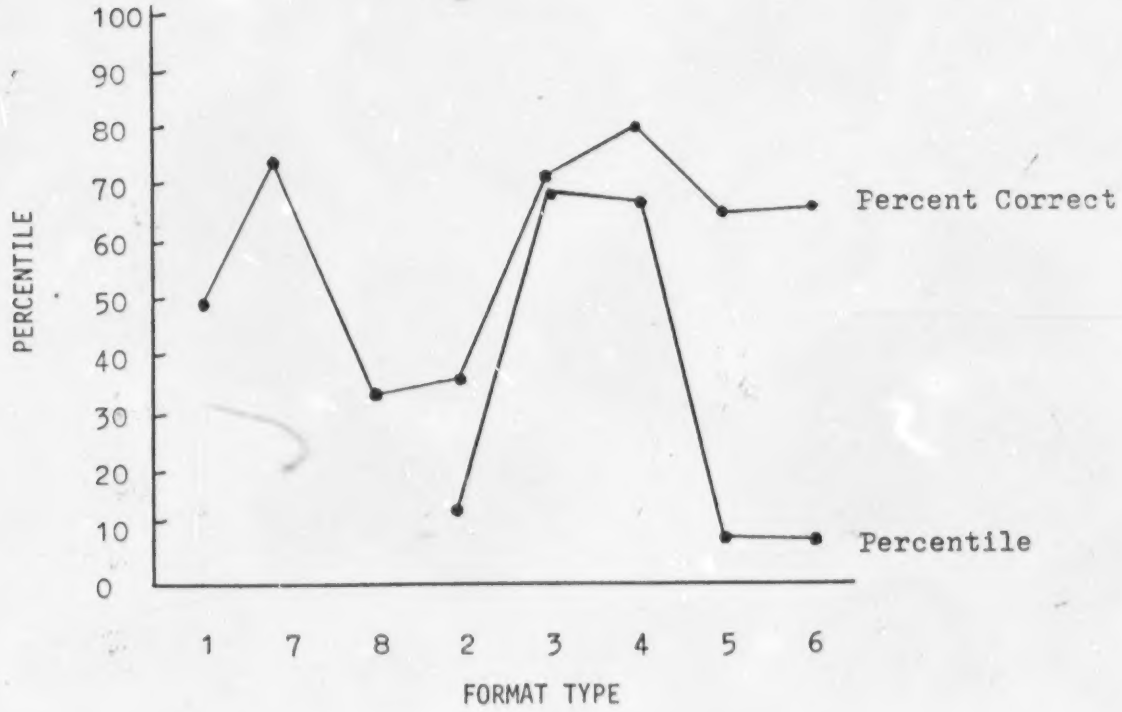
[b]Standard deviation.

Table 7

Sources of Variance for Repeated Measures

ANOVA for Sessions and Classes

Source	Degrees of freedom	Error term	Sum of squares	F
Sessions (S)	1	R(S)	68.45	.726
Classes (C)	1	CR(S)	16.2	.206
Replicates R(S)	38		3582.75	
S X C	1	CR(S)	14.45	.184
CR(S)	38		2984.35	



Note. Percentile rankings cannot be ascertained for formats 1, 7, and 8.

Figure 1. Differences in percentile rank between different formats.